No More 404s: Predicting Referenced Link Rot in Scholarly Articles for Pro-Active Archiving

No More 404s: Predicting Referenced Link Rot in Scholarly Articles for Pro-Active Archiving

Ke Zhou
University of Edinburgh
Ke.Zhou@ed.ac.uk

Claire Grover
University of Edinburgh
Claire.Grover@ed.ac.uk

Martin Klein
University of California Los Angeles
martinklein@library.ucla.edu

Richard Tobin
University of Edinburgh
richard@inf.ed.ac.uk

ABSTRACT

The citation of resources is a fundamental part of scholarly discourse. Due to the popularity of the web, there is an increasing trend for scholarly articles to reference web resources (e.g. software, data). However, due to the dynamic nature of the web, the referenced links may become inaccessible ('rotten') sometime after publication, returning a "404 Not Found" HTTP error. In this paper we first present some preliminary findings of a study of the persistence and availability of web resources referenced from papers in a large-scale scholarly repository. We reaffirm previous research that link rot is a serious problem in the scholarly world and that current web archives do not always preserve all rotten links. Therefore, a more pro-active archival solution needs to be developed to further preserve web content referenced in scholarly articles. To this end, we propose to apply machine learning techniques to train a link rot predictor for use by an archival framework to prioritise pro-active archiving of links that are more likely to be rotten. We demonstrate that we can obtain a fairly high link rot prediction AUC (0.72) with only a small set of features. By simulation, we also show that our prediction framework is more effective than current web archives for preserving links that are likely to be rotten. This work has a potential impact for the scholarly world where publishers can utilise this framework to prioritise the archiving of links for digital preservation, especially when there is a large quantity of links to be archived.

Categories and Subject Descriptors: H.5.4 [Information Interfaces and Presentation: Hypertext][Architectures, Navigation]
Keywords: Digital Preservation, Repositories, Web Persistence

1. INTRODUCTION

The citation of resources is a fundamental part of scholarly discourse. Beyond traditionally-cited published articles or books, in the digital age web-based scholarly endeavour has greatly enlarged the range of scholarly artefacts that are being published and referenced. Many of these are resources created as part of research activity such as software, datasets, presentations, videos, etc. as well as scientific workflows and ontologies. Our recent research[1] [11] has shown that in large-scale scholarly corpora, around 20% of scholarly articles have referenced web links (URLs) and the number of those referenced links is increasing over the years.

The real-time nature of the web enables immediate access to web resources and dramatically increases the speed of knowledge dissemination. At the same time however, it also poses the challenge of preserving endangered referenced web links that may become inaccessible (i.e. rotten) after publication. There are two ways in which links can be considered dysfunctional: (1) *link rot*, the content of the link is not available on the live web at its original URI anymore; or (2) *content drift*, the content of the link has changed since publication of the scholarly article. Both scenarios give rise to the risk that researchers in the future will not be able to thoroughly study the citation context of a scholarly publication. In this preliminary work, we focus solely on investigating the first type of link rot problem, i.e. the links returning a "404 Not Found" HTTP error status code. We leave the second type of reference rot problem, content drift, for future work.

In previous work, various researchers have aimed to quantify aspects of the reference rot problem [4, 7, 6]. There exist a variety of web archival services [1] (e.g. Internet Archive[2]) which aim to preserve web resources. Although these archival services largely preserve online resources [1], how well they archive referenced web resources for the scholarly world is not clear. Klein et al. [6] recently investigated this problem and found that for several large-scale scholarly collections, one out of five STM (Science, Technology, and Medicine) articles suffer from reference rot, meaning it is impossible to revisit the entire web context that surrounds them at some point after their publication. Therefore, more pro-active archival solutions for archiving links in the scholarly world are required for digital preservation for future researchers. Given the large and increasing number of links referenced in the scholarly world, we might not have the capability to archive all of them and we therefore need to prioritise some links over others. We hypothesise that *a referenced link that is more likely to become rotten should become a higher priority to be pro-actively archived*. Even if all of the links can be archived, it would still be useful to automatically suggest to publishers or authors the links that are more likely to become rotten so that action can be taken.

Our main goal in this work is to investigate whether it is possible to accurately predict link rot. We aim to answer the following research question: **Can we accurately predict referenced link rot in scholarly articles in order to prioritise pro-active archiving**

[1]The Hiberlink project (http://www.hiberlink.org/) is supported by the Andrew W. Mellon Foundation. We would like to thank our project partners from EDINA and Los Alamos National Laboratory Research Library for their useful feedback.
[2]https://www.archive.org/

of links that are at risk?

The contributions of this paper are two-fold: **(1)** We demonstrate the feasibility of using a machine learned classification framework to accurately predict the referenced link rot problem (i.e. the likelihood that a given link will become rotten). We also analyse the impact of different features (including scholarly article features and link features) on the link rot prediction task. **(2)** In order to demonstrate the effectiveness of our link rot predictor, we simulate pro-active archiving and show that the approach outperforms current web archives in preserving rotten links.

2. RELATED WORK

Two lines of research relate to this work. One focuses on current endeavours to study and quantify the referenced link rot problem in the scholarly world. The second line focuses on reviewing current web archives and their archiving (crawling) strategies. The contributions of our work lie in our proposed machine learned link rot prediction framework and extensive analysis of the features that affect the link rot problem in the scholarly world.

Referenced Link Rot Study Over the past ten years, extensive, although typically small-scale, research has been conducted on the persistence of the resources identified by URLs cited in scholarly publications, especially journal articles. For example, the study by Lawrence [7] was seminal and indicated that only 75% of URLs were accessible in the corpus of citations he examined in 2000. It also attempted to discover if the resources that were not available at their original URL were still online at new locations. For example, Lawrence [7] used search engines to investigate the availability of 205 URLs that did not resolve and rediscovered 163 (79.7%). The studies since have all been small-scale while the extent to which the cited resources were available from archives was only comprehensively studied recently [6]. By using the Memento protocol [10] (a web archive aggregator), this research reported that the survival rate in larger STM (Science, Technology, and Medicine) corpora is at around 80%. A more detailed review of those studies can be found in [6] and our work is inspired by those studies.

Web Archive and Crawling Archiving web content is not new and there exists a variety of web archives [1] and their aggregators [10]. The way they prioritise preservation is based on several heuristics [5, 9], similar to the way in which search engines crawl web pages [2]. Different features, mostly derived from web pages and their domain, such as PageRank, etc. have been exploited. For example, current web archives [5] prioritise archiving of web pages from top ranked domains (although with only a limited number of levels). Crawling the web pages of one site at a time can be done in breadth first mode, postponing the crawling of external web pages until the corresponding sites are visited.

Unlike current web archives, the links of interest to us for preservation are referenced links within scholarly articles. As we briefly show in Section 3.1, current web archives fail to preserve all the referenced link contents for publishers and future researchers. We aim to use the features derived from both the referenced link and scholarly articles to predict link rot. This prediction is then used to assist a pro-active archive to prioritise the archiving of referenced links that are more likely to become rotten and that should therefore be preserved as early as possible.

3. LINK ROT PREDICTION

In this study, we treat link rot prediction as a binary classification problem (i.e. we classify links as "highly likely to become rotten" or not), and investigate how to use machine learning techniques to learn this. In this section, we first quantify the link rot problem, followed by presentation of our approach and evaluation results.

Table 1: The characteristics and quantification of referenced link rot of the Elsevier scholarly article collection.

Statistics and Results/Collection	Elsevier [3]
(a). Subject	variety of subjects, e.g. finance, medical
(b). Publication Type	Journal and Book Series
(c). Publication Period	1997-2012
(d). # of articles	648,388
(e). fraction of articles with links	12.1% (78,237)
(f). total # of links extracted	193,955
(g). fraction of "rotten" links	36.2% (70,270)
(h). fraction of archived links	77.5% (150,368)
(i). fraction of archived "rotten" links	62.3% (43,745)

3.1 Link Rot Quantification

By using a state-of-the-art link extraction system [11] (with high performance F-measure of 0.8) on a scholarly collection, we aim to quantify the referenced link rot problem by dereferencing the links on the live web and obtaining their archived status via Memento (a web archive aggregator). The aim here is to introduce link rot quantification (following approaches similar to previous work [6]), leaving a more thorough quantification for future work.

We use the Elsevier collection as our test set—detailed characteristics of this collection are shown in Table 1(a) to (d). We can observe that this collection is relatively large, with hundreds of thousands of scholarly articles on a variety of subjects spanning more than fifteen years. From Table 1(e) to (f), we can also see that a significant fraction of documents (12%) have referenced web links.

To quantify link rot, we probe each extracted referenced link on the live web and check its HTTP status. Since there could potentially be redirects of the links, we set a rule to allow redirects only up to a maximum of 50. We record the whole HTTP transaction chain and if this ends with a 2XX status code, we consider the link to exist (i.e. it is not rotten). Otherwise, we consider the link to be rotten. The results are shown in Table 1(g). We can observe that many links are rotten and 36.2% of the links extracted suffer from the risk of content rot. This finding reaffirms previous research [6]. Not surprisingly, we also find this problem occurs across all publication time spans and subjects. We conclude it is crucial for a digital preservation service to preserve all those referenced web links.

Using a Memento Aggregator [10] that covers nine archives, including the Internet Archive, Web Citation, the UK National Archive and the Library of Congress, we also attempt to quantify whether the links have been archived by current web archival facilities. Specifically, we retrieve a TimeMap for each of the referenced URLs. If a TimeMap cannot be retrieved, the URL is marked as not being archived, and otherwise marked as being archived. We conducted this study in March 2014 and Table 1 (h) and (i) present the results. From (h), it can be seen that for all the links extracted (Table 1 (f)), a large percentage of them (77.5%) is archived at least once over the years. However, from (i), we can observe that for the links that are rotten (Table 1 (g)) only 62.3% of them are preserved by current web archives. This implies that the remaining approximately 40% of the rotten link contents are not preserved and would not be retrievable by future researchers. We conclude, therefore, that the current digital preservation framework fails to accurately preserve all of the referenced link content in scholarly works and that a more pro-active referenced link preservation framework needs to be developed.

[3] http://www.developers.elsevier.com/cms/index

Table 2: Two types of features generated for quantifying the likelihood of link rot for machine learning in the scholarly world.

Feature	Description	Data Source
Scholarly Article Features		
Publication Subject Vector	A vector containing all the subject areas[4] of the publication where the weight of the given publication's subject is 1, otherwise 0.	meta-data
Year of Publication (distance to the present)	An integer score representing the distance (in years) from the publication year of the given publication to 2014.	meta-data
Open Access Status of the Publication	Whether the publication is open-access (1) or not (0).	meta-data
h5-index of the Journal (if available)	The h5-index of the given publication evenly distributed into 20 quality-bins.	Google Scholar Metrics[5]
Link Features		
Link Domain PageRank	The pagerank score of the given link's score, averaged over its domain based on pagerank values computed on 50 million web pages then evenly distributed into 100 quality-bins.	ClueWeb'09 dataset[6]
Link Depth	An integer score of the depths of the link (i.e. number of tokens).	URL standard tokenization[7]
Link Position Vector	A vector containing all the link positions where the weight of the given link position is 1, otherwise 0.	XML annotation
Link Type Vector	A vector containing all the link types where the weight of the given link type is 1, otherwise 0.	ODP and UClassifier[8]

3.2 Link Rot Prediction Approach

Given the severity of link rot in the referenced links in scholarly works, we aim to predict the likelihood of this link rot in order to preserve the links that are highly likely to become rotten in a pro-active archival framework. In order to train and test a machine learned link rot predictor, we need to develop various features to effectively represent the problem and train the classifier.

3.2.1 Features

We believe that two types of factors influence the likelihood of the link rot: (1) scholarly article features: the quality and the type of the scholarly publication which the link is extracted from; and (2) link features: the quality and the type of the link. The underlying hypothesis is that some of the links (e.g. from high-quality domains) originating from some scholarly publications (e.g. open-access journals) might be less susceptible to rot than others. Details of the features we extract to quantify this are presented in Table 2. The main objective of this preliminary study is to demonstrate the feasibility of our approach.

Most of the features are either available from the Elsevier collection meta-data (e.g. publication year) or from existing resources (e.g. h5-index, PageRank). To obtain the link position in the article, we extract the annotated sections within the XML format of the Elsevier collection using a stylesheet. We define a set of manual rules to transform the XML annotations to the set of document structures we are interested in. The corresponding positions we obtain are: header, footnote, figure, table, body and reference.

We also believe that link type information (specifying which resource a link refers to) could also be useful for link rot prediction. To investigate link type, we use a publicly available classification tool, UClassifier, to classify links into the taxonomy provided from the Open Directory Project (ODP)[7]. This is a machine learning classifier that is based on training data provided by ODP (textual representation of all the web pages within each category). Rather than being interested in the general topics of the link (e.g. Arts, Business, Computers, etc.), we are more interested in whether the links appear within categories that are more related to scholarly publications: software, licence, data, slides, blog, image, video and publishers. We manually label the corresponding sub-categories in the ODP with those categories. All the links that are not classified into any of the categories are given the category "Other". We also have a whitelist of publishers' website domains to determine whether a link points to a publisher website. To represent each link for learning, rather than downloading the actual content (which might be not available due to the "rot" problem), we use the textual context of each referenced link for the representation. Following one of the best performing methods from Ritchie [8] for representing citation context, we use "three sentences" around a referenced link as the textual context. Although a more comprehensive evaluation of this link type classifier would help us better estimate its effectiveness, the idea in this work is to apply current solutions. Full evaluation of our link type classifier is left for future work. We empirically demonstrate in Sec. 3.3 that this feature is useful for link rot prediction.

3.2.2 Classifier and Training

We use Support Vector Machine (SVM) learning for our classifier since SVMs are proven to perform well in other classification tasks. Specifically, we use the publicly available LIBSVM toolkit (`http://www.csie.ntu.edu.tw/~cjlin/libsvm/`) for our implementation. Study of the effectiveness of other classifiers (e.g. linear regression, random forest, etc.) is left for future work.

To train the classifier, we sample $10,000$ links from the Elsevier collection and obtain the corresponding rot label (1 or 0) by probing on the live web. To avoid bias where the classifier rewards a class with more positive cases (here, the non-rotten links), our sampling approach is based on random sampling while ensuring that the two classes contain the same number of positive cases. In order to train the performance of our classifier, we perform five-fold cross validation with the sampled links using our approach. We finally report the standard AUC [3] of our trained classifier on the test set (another sampled set of 2,000 links using the same sampling approach) and we ensure that the test set does not overlap with the training set.

3.3 Evaluation

We conduct two types of evaluation: (1) the AUC of the learned link rot predictor; and (2) the effectiveness of applying the link rot predictor to a simulated archival environment.

3.3.1 Evaluating the Link Rot Predictor

The evaluation results of our link rot predictor are presented in Table 3. Specifically, we find that we can obtain an AUC of prediction up to 0.72, which is significantly better than a random prediction (0.50 AUC). Significance was tested using a sign test, where the null hypothesis is that the classifier predicts the link rot randomly with equal probability. This demonstrates that our proposed learned approach and corresponding features are feasible and effective in predicting link rot.

To further investigate the effectiveness of each feature in its contribution to the prediction, we conduct an ablation study (i.e. leave one feature type out and track the performance change). The results are shown in Table 3. A non-significant performance drop in AUC does not necessarily mean the feature captures no useful evidence, as features may be correlated. We can observe the following

[4] Elsevier contains 27 subjects from http://www.elsevier.com/journals/title/a

[5] http://scholar.google.co.uk/citations?view_op=top_venues

[6] http://boston.lti.cs.cmu.edu/clueweb09/wiki/tiki-index.php?page=PageRank

[7] http://www.ietf.org/rfc/rfc1738.txt

[8] UClassifier (http://www.uclassify.com/browse) and ODP (http://www.dmoz.org)

Table 3: Feature set contribution to link rot prediction AUC: leaving one feature type out (feature ablation study). The differences of AUC performance are calculated over "All", our classifier using all features.

Feature Variation	Feature Type	AUC	% diff
All	Both	**0.72**	
no.Publication Subject	Article	0.70	-2.8%
no.Year of Publication	Article	0.65	**-9.7%**
no.Publication Open Access Status	Article	0.72	-0.0%
no.h5-index of the Journal	Article	0.70	-2.8%
no.Link Domain PageRank	Link	0.67	**-6.9%**
no.Link.Depth	Link	0.66	**-8.3%**
no.Link Position	Link	0.71	-1.4%
no.Link Type	Link	0.69	**-4.2%**

trends: (1) In terms of feature types, in general, more link features contribute more significantly than scholarly article features. Specifically, the article feature "Year of Publication" contributes most to the prediction AUC. To explore this further, we plot the rot likelihood of extracted links according to publication year in Figure 1. We can observe that links are more likely to be rotten if they originate from older scholarly publications. Not surprisingly, the further from the time of publication, the more likely it is that an extracted link will be rotten. Note that the link creation time might potentially be correlated with the year of publication since the links are likely created prior to the scholarly articles which cite them. (2) The second and third most helpful features are link features: "Link.Depth" and "Link.Domain.PageRank". From close examination, we find that links to low quality domains and which have more depth (longer length) are more likely to rot. (3) The features which contribute least are link position and publication open access status. This implies that ease of access and the position of a link in a scholarly article do not have a big impact on link rot. It is also interesting however that the citation-based quality measure of the publication (h5-index) can also contribute to link rot prediction. This implies that links to scholarly articles, which are published in higher impact journals, are less likely to be rotten.

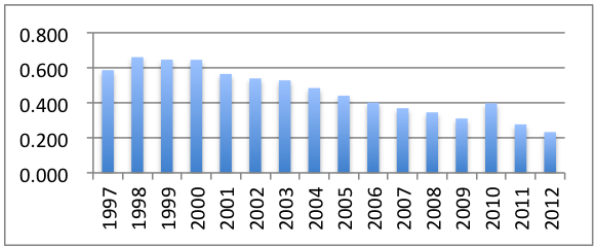

Figure 1: Time-aware analysis of likelihood of extracted links deemed to be rotten according to publication year (Elsevier).

In summary, we have analysed different features in predicting link rot and have demonstrated the feasibility of our approach.

3.3.2 Evaluating Simulated Pro-active Archiving

So far, we have demonstrated the effectiveness of predicting the likelihood of link rot. However, the utility of this learned link rot predictor in a pro-active archive for scholarly works is still to be considered. Since there are currently no pro-active archive solutions for scholarly works, we simulate one using our link rot predictor and compare its effectiveness in archiving rotten links from past publications, compared to current web archives.

We first sample a set of extracted links from all the links extracted from Elsevier (Table 1). Then we use our learned link rot predictor to predict which links are more likely to become rotten and should be prioritised for archiving. Finally, by archiving the same number of links with the current web archives, we track whether our simulated pro-active archive can preserve more links at risk of rot. The results are presented in Table 4. We can observe that compared with current archives (62.3%), our simulated pro-active archive preserved a significantly larger number of rotten links (84.8%). Although this simulation does not conform to a real-world setting, we conclude that our link rot predictor can be helpful for archiving scholarly referenced links for digital preservation.

Table 4: Evaluation of applying link rot predictor to archive rotten links, compared with current archival solution.

Results /Systems	Current Archives	Simulated Archive
fraction of archived "rotten" links	62.3%	84.8%

4. CONCLUSIONS

In this paper, we briefly quantified the referenced link rot problem in scholarly works and proposed a link rot prediction task. We reaffirm previous research that link rot is prevalent in the scholarly world and that pro-active archival solutions are required to solve this. To this end, we proposed a machine learned link rot predictor and investigated two sources of evidence for learning, i.e. link features and scholarly article features. We treat link rot prediction as a binary classification problem and use machine learning techniques (SVM) to learn a classifier. Although we have only used a limited set of essential features, we found that we can predict the link rot problem with an AUC of 0.72. We also showed that this preliminary predictor could be used alongside an archival framework to prioritise pro-actively archiving links that are more at risk of rot. We tested this in a simulated environment and showed that the simulated archival solution outperforms current web archives. Although testing this in a real-world archival setting is out of scope for this paper, we believe that with further feature engineering and classifier evaluation, this could be further improved and used in a practical setting.

Our future work includes expanding our link rot definition and investigating the more complex *content drift* aspect of the reference rot problem: the change of referenced link content over time. We would also like to study more features and conduct a more thorough evaluation of our learned link rot predictor for assisting scholarly referenced link archiving.

5. REFERENCES

[1] S. G. Ainsworth, A. Alsum, H. SalahEldeen, M. C. Weigle, and M. L. Nelson. How much of the web is archived? In *JCDL*, JCDL '11, pages 133–136, 2011.

[2] J. Cho and H. Garcia-Molina. The evolution of the web and implications for an incremental crawler. In *VLDB*, pages 200–209, 2000.

[3] T. Fawcett. An introduction to ROC analysis. *Pattern recognition letters*, 27(8):861–874, 2006.

[4] D. Fetterly, M. Manasse, M. Najork, and J. Wiener. A large-scale study of the evolution of web pages. In *WWW*, pages 669–678. ACM, 2003.

[5] D. Gomes, S. Freitas, and M. J. Silva. Design and selection criteria for a national web archive. In *Research and Advanced Technology for Digital Libraries*, pages 196–207. Springer, 2006.

[6] M. Klein, H. Van de Sompel, R. Sanderson, H. Shankar, L. Balakireva, K. Zhou, and R. Tobin. Scholarly context not found: One in five articles suffers from reference rot. *PloS one*, 9(12):e115253, 2014.

[7] S. Lawrence, D. M. Pennock, G. W. Flake, R. Krovetz, F. M. Coetzee, E. Glover, F. Å. Nielsen, A. Kruger, and C. L. Giles. Persistence of web references in scientific research. *Computer*, 34(2):26–31, 2001.

[8] A. Ritchie, S. Robertson, and S. Teufel. Comparing citation contexts for information retrieval. In *CIKM*, pages 213–222. ACM, 2008.

[9] M. Spaniol, D. Denev, A. Mazeika, G. Weikum, and P. Senellart. Data quality in web archiving. In *Proceedings of the 3rd workshop on Information credibility on the web*, pages 19–26. ACM, 2009.

[10] H. Van de Sompel, M. Nelson, and R. Sanderson. HTTP framework for time-based access to resource states–Memento, 2012. http://tools.ietf.org/html/rfc7089.

[11] K. Zhou, R. Tobin, and C. Grover. Extraction and analysis of referenced web links in large-scale scholarly articles. In *Proceedings of the 14th ACM/IEEE-CS Joint Conference on Digital Libraries*, JCDL '14, pages 451–452, 2014.